WHITE PAPER
USING BIBLIOMETRICS IN EVALUATING RESEARCH

DAVID A. PENDLEBURY
RESEARCH DEPARTMENT, THOMSON REUTERS, PHILADELPHIA, PA USA

SCIENTIFIC

THOMSON REUTERS

INTRODUCTION:
THE MAIN TOOL OF SCIENCE

Counting, measuring, comparing quantities, analyzing measurements: quantitative analysis, as Lord Kelvin's famous observation suggests, is perhaps the main tool of science. In this century, the volume of scientific research—measuring to "know something"—and recording and communicating that knowledge through publications, has itself become enormous and complex. Science research is now such a large enterprise and the substance of scientific research is so complex and specialized that personal knowledge and experience are no longer sufficient tools for understanding trends or for making decisions.

"IF YOU CAN MEASURE THAT OF WHICH YOU SPEAK, AND CAN EXPRESS IT BY A NUMBER, YOU KNOW SOMETHING OF YOUR SUBJECT; BUT IF YOU CANNOT MEASURE IT, YOUR KNOWLEDGE IS MEAGER AND UNSATISFACTORY."

William Thomson, Lord Kelvin

Yet, the need to be selective, to highlight significant or promising areas of research, and to manage better investments in science is only increasing. Around the world, the purses of government, industry, education and philanthropy have not grown as fast as science. Those in universities, government offices and labs, and boardrooms face pressing questions about what should be supported and what should not, or which research projects and researchers should receive more support than others.

Achievements in industry can often be tracked directly by counting patents and measuring, in terms of sales revenues, the commercial success of discoveries as they move from the laboratory to the marketplace. However, the achievements and trends in science are not so easily counted. And so, until relatively recently, peer review has been the main route by which science policymakers and research funders have coped with decisions on what course to set for science.

However, the ever-increasing size and specialized nature of research today makes it difficult for a small group of experts to evaluate fully and fairly the bewildering array of research, both that accomplished and that proposed. A library faced with collection decisions, a foundation making funding choices, or a government office weighing national research needs must rely on expert analysis of scientific research performance.

Peer review still represents the standard approach to research evaluation and decisions about allocating resources for science. Experts reviewing the work of their colleagues should rightly be the basis of research evaluation. However, it should be one of several approaches to making decisions. For the inevitable bias in peer review, whether intentional or inadvertent, is widely recognized as a confounding factor in efforts to judge the quality of research.

BIBLIOMETRICS

One such approach is bibliometrics (sometimes called scientometrics). Bibliometrics turns the main tool of science, quantitative analysis, on itself. At its most fundamental, this approach to research evaluation is simply counting. The complexity is in the analysis and use of the numbers, for the statistics obtained can be understood as indicators of achievement or lack thereof.

Quantitative evaluation of publication and citation data is now used in almost all nations around the globe with a sizeable science enterprise. Bibliometrics is used in research performance evaluation, especially in university and government labs, and also by policymakers, research directors and administrators, information specialists and librarians, and researchers themselves.

The Development of Publication and Citation Analysis

There are, of course, many activities and outcomes of research that can be counted. Perhaps the most basic and common is number of publications, which may be used as a measure of output. Citations, the references researchers append to their papers to show explicitly earlier work on which they have depended to conduct their own investigations, shows how others

use a work in subsequent research. Tracking citations and understanding their trends in context is a key to evaluating the impact and influence of research.

A citation index for science was first described in 1955 by Eugene Garfield, the founder and chairman emeritus of what was then ISI, in the journal *Science*. He realized his vision a few years later with the production of the 1961 *Science Citation Index®*.

The original and continuing main purpose of Garfield's citation database is improved or expanded information retrieval. By recording not only bibliographic information on the journal articles covered but also the cited references in these journal articles, Dr. Garfield offered researchers a way to find articles relevant to their work that they would not otherwise turn up by searching author names, title words, or subject headings alone.

The operating principle of a citation index is this: If a researcher knows of a publication important to his or her work, a citation index would allow the researcher to identify journal articles published subsequent to that work that cited it. On the assumption that the citing journal article is related in some way to the substance of the cited work, the user of the citation index may, as it were, search forward in time to uncover studies of interest potentially profitable for the researcher's work.

It did not escape Garfield's notice that such a database could serve other purposes as well, such as monitoring and analyzing the structure and growth of science. Others, too, saw this possibility. Among them are the historian of science Derek J. de Solla Price, author of the 1963 classic *Little Science, Big Science*, and the sociologists of science Robert K. Merton, Jonathan and Stephen Cole, Warren Hagstrom, and Diana Crane. Francis Narin of CHI Research in the United States was also an early proponent and pioneer in using ISI publication and citation data to analyze science, particularly through his influential *Evaluative Bibliometrics* of 1976.

The combination of an ever-growing corpus of publication and citation data compiled by ISI over the 1960s and 1970s and the simultaneous increase in computing power and software applications, especially those developed in the 1980s and 1990s, has made bibliometrics a practical and even cost-effective pursuit.

Bibliometrics and Peer Judgment: A Two-Pronged Approach

Never in its long history has ISI, now Thomson Reuters, advocated that bibliometrics supercede or replace peer judgements. Rather, publication and citation analysis is meant to be a supplement to peer review. The two together—peer review and quantitative analysis of research—better inform evaluation and decisions. For quantitative analysis offers certain advantages in gathering the objective information necessary to decision-making:

- Quantitative analysis of research is global in perspective, a "top-down" review that puts the work in context, complementing the local perspective of peer review. Quantitative research analysis provides data on all activity in an area, summaries of these data, and a comprehensive perspective on activity and achievements.
- Weighted quantitative measures, such as papers per researcher or citations per paper, remove characteristics, such as the place of production, or past reputation, that color human perceptions of quality.

  For example, when we think of "the best," it is hard not to think automatically of the biggest producers, such as individuals, labs, and universities; but those locations might not be the source of the most important work. Likewise, quantitative research analysis indicates current top performers, thus balancing human perceptions of reputation.

The Growing Use of Bibliometrics

For these reasons and others, nations with significant science enterprises have embraced bibliometrics. Today, bibliometrics programs with large teams of analysts are firmly established in many nations, and these groups issue bibliometric reports, often called science indicators studies, at regular intervals.

A few such groups are the National Science Foundation (United States); the European Commission; and Japan's National Institute for Informatics (NII), National Institute for Science and Technology Policy (NISTEP), and Ministry of Economy, Trade and Industry (METI). Other nations with active bibliometrics groups include Australia, Belgium, Brazil, Chile, China, Finland, France, Germany, Israel, Italy, The Netherlands, Norway, Portugal, South Africa, South Korea, Spain, Sweden, Switzerland, and Taiwan. In almost all cases, the publication and citation data of Thomson Reuters form the basis of their bibliometric analyses.

TEN RULES FOR USING PUBLICATION AND CITATION ANALYSIS

In light of science's ever-growing complexity and the challenge of rationing limited resources, government policymakers, managers, and others turn to quantitative analysis of research to help make their task easier.

However, employing quantitative indicators of research performance adds an extra burden to decision-makers: they must strive to define clearly the data they need and work to understand the significance of the analyses. This extra effort is necessary and worthwhile, both for the greater understanding and practical help the data offer, as well as for the beneficial effect of the data: adding fairness to evaluation and helping to prevent abuses that may arise from small-scale, closed peer review.

It is important to understand what quantitative research analysis offers evaluators and decision makers—and what it cannot possibly deliver. For these techniques or tools can never be a substitute for human judgment. It is important that measurement and judgment be used in tandem, and that the quantitative tools are used appropriately.

Numbers alone can be dangerous because they have the appearance of being authoritative. In the face of statistics, many discussions stop. And that is unfortunate: numbers instead should fuel discussions and help illuminate features in the research landscape that might otherwise be overlooked. And when the purpose of pursuing quantitative analysis of research is for "window dressing" or to prove to policymakers, administrators, or funding agencies something decided

upon even before the data are collected and analyzed, such effort works against the true goal of the analysis.

To that end, Thomson Reuters offers some practical advice on how to approach and evaluate research performance using quantitative indicators, such as those we and others make available.

Ten practical rules—a kind of checklist—may prove helpful to those faced with a need to analyze research performance using publication and citation data. Although neither canonical nor exhaustive, these are useful rules of thumb.

1) Consider whether available data can address question.

Before even beginning an analysis, ask if the data available are sufficient to analyze the research under review. A general observation: The more basic the research and the larger the dataset, the more likely that the analysis will be reliable at face value. This implies, of course, that the more applied the research and the smaller the dataset, the more likely the picture obtained will be partial, contain artifacts, and therefore possibly lead to misinterpretation.

2) Choose publication types, field definitions, and years of data.

These three technical details are critical in defining what body of research is to be analyzed.

Publication types. The standard practice is to use journal items that have been coded as regular discovery accounts, brief communications (notes), and review articles—in other words, those types of papers that contain substantive scientific information.

Traditionally left to the side are meeting abstracts (generally not much cited), letters to the editor (often expressions of opinion), and correction notices.

Field definitions. Categorization is always a difficult problem. While a researcher may consider himself or herself to be an immunologist, a theoretical physicist, or a soil scientist, the papers may appear in journals not typically assigned to the self-described field. This becomes an issue when one wishes to use relative measures—to compare their work and citation averages with those of their field. The reality is that no one and nobody's work fits perfectly into a single category. In fact, each researcher publication list is a unique record.

Time frame. Finally, one must decide which years of publications and of citations to use. These do not have to be the same. For example, one might want to review the last 10 years of papers but only the last five years of citations to these papers. At other times, the publication and citation window might be identical and overlapping.

Generally, when citations are to be used to gauge research impact, Thomson Reuters recommends at least five years of publications and citations, since citations take some time to accrue to papers. In the fastest moving fields, such as molecular biology and genetics, this might take 18 months to two years, whereas in others, such as physiology or analytical chemistry, the time lag in citations might be, on average, three, four, or even five years.

3) Decide on whole or fractional counting.

It is important to decide which approach, whole or fractional counting of output, best serves the evaluation purposes.

This would seem a simple task were it not for the difficulty of determining authorship and institutional sponsorship. Multi-authored publications, increasingly the norm in science, raise an important technical question for those trying to understand productivity and influence.

Researchers rarely distinguish who is responsible for how much of the work is reported. Often no lead author is indicated, and when one is, in some fields the first name listed is traditionally the lead, but in other fields it is the last name listed. Even when a lead author is indicated, there is never a quantitative accounting of credit. In effect, the presentation suggests that all are equal in their authorship and contributions, although this cannot be possible. Level and nature of contributions vary and in some cases there is honorary authorship.

Without a way to identify contribution, should each author or institution listed on a paper receive whole or a fractional—a proportionate—share of the paper and, for that matter, the citations that it attracts?

Consider a paper by three scientists at three different universities that has been cited 30 times. Should each receive credit for one-third of the paper and, say, 10 citations (one-third of the citations)? Or should each receive a whole publication count and credit for all 30 citations? Another possibility would be to use fractional publication counts but whole citation counts. Measuring this way, each researcher or university would receive credit for one-third of the paper but also for all 30 citations.

On the belief that anyone appearing as an author should be able to explain and fully defend the contents of the paper, Thomson Reuters almost always uses whole publication and citation counts. But career concerns and the simple approach of using the length of one's list of publications as a measure of achievement has brought about much unwarranted authorship. Two fields, in particular, may demand fractional counting: high-energy physics and large-scale clinical trials. These fields produce papers listing hundreds of authors and almost as many institutions. Faced with such papers, one must ask if these reports are really attributable to any scientists or any institutions.

Thomson Reuters records all authors and addresses listed on a paper, so these papers can be attributed to all producers. This is not always the case with other databases, many of which provide only the first name listed or reprint author and his or her address. Thomson Reuters' comprehensive indexing of authors and institutions is one reason why the citation database is the preferred source for bibliometric analyses (the others are its multidisciplinary coverage and, of course, citations, which can be used as measures of influence and impact).

4) Judge whether data require editing to remove artifacts.

Artifacts are aspects of the data that may confound the analysis or mislead the analyst. For example, in a small

to medium-sized dataset, the inclusion of papers with hundreds of author names or institutional addresses, counted whole, can have this effect. A very basic kind of artifact — variation — is extremely important and requires careful editing.

Artifacts of variation. The various ways a name might appear can also confound the analysis. For example, authors are not uniform in how they list their own institution, so institution names often appear in varying forms. Likewise, author names can appear in varying forms: on some papers, the name might be listed as PENDLEBURY D, on others, as PENDLEBURY DA, with the effect of seeming to be two different people. Conversely, the papers of many different people who share the same name form (such as SUZUKI T or LEE K), could be lumped together as if representing the work of one author.

To remove the artifacts, each paper must be manually reviewed to determine correct authorship and unify variant designations under one, preferred name. This is dreary but important work. Thomson Reuters records the institutional name as the one given on the original paper, but the variant designations are unified so that the statistics for an institution appear under one, preferred name.

Other possible artifacts. Three common objections to the use of citation counts in evaluating research have to do with possible artifacts in the data: negative citations, the "over citation" of review articles and methods papers, and self-citation or citation circles. Although all could be a conflating factor in very small datasets, generally these concerns have not been shown to significantly distort citation counts.

1) Negative citations are few in number. They are rare events, statistically speaking. Scientists typically cite for neutral or positive reasons—to note earlier work or to agree with and build upon it. Assessing whether a citation is positive or negative requires a careful, informed reading of the original paper, and this obviously cannot be attempted with more than a few hundred papers at the most.

Of several articles published in which this sort of analysis was attempted, always dealing with a sample of papers in a single field that could be controlled by the analyst, outright negative citations were few, on the order of 5 percent or less. Naturally, we all recall notorious examples, such as the case of the Cold Fusion paper, but these and other rare cases are the so-called exceptions that prove the rule. Frequency of citation, many studies have shown, correlates positively with peer esteem. Negative citations, to the degree they appear, are little more than background noise and do not materially affect our analyses.

2) Are methods papers and reviews "over cited?" A method may be used by many and acknowledged often, if perfunctorily. A review offers a convenient and concise way of recognizing and summarizing the previous literature in an area, and thus becomes a convenient reference. But it stands to reason that only useful methods and only good reviews are highly cited.

Methods papers and reviews are seen by some as lesser contributions than discovery accounts. Consider, however, the paper that did more than any other to accelerate research and discoveries in molecular biology and genetics in the 1980s and 1990s, the Kary B. Mullis paper describing the polymerase chain reaction technique, work for which Mullis won the Nobel Prize in Chemistry in 1993. So, too, reviews synthesize disparate work and can have great catalyzing effects in moving fields forward. However, if these types of publications are a concern for those using the analysis, they can be removed.

3) Questions are often raised about self-citation and citation circles or cabals, in which a group of researchers agree to cite one another to boost their citation totals. Self-citation is a normal and normative feature of publication, with 25 percent self-citation not uncommon or inordinate in the biomedical literature, according to several studies. It is only natural that when a researcher works on a specific problem for some time, he or she would cite earlier publications from this effort. A medicinal plant biologist, for example, who exhibited 75 percent self-citation in her papers, might be virtually the only person working in this specific area.

Someone determined to boost citation counts through unbridled self-citation would need to overcome several obstacles. The first is peer review—objections from reviewers and the journal editor to unnecessary citations and perhaps the absence of citations to other appropriate work. A researcher might aim to publish in lower-impact journals with looser standards of review, but as fewer people would see and cite these articles, the researcher would lose some degree of opportunity for citations from others. It seems a self-defeating strategy. Citation circles call to mind the mythical unicorn: everyone can imagine it, but none have so far been produced.

5) Compare like with like.
Those using quantitative evaluation of research should keep in mind that the methodology for a bibliometric analysis always compares like with like, "apples with apples," not "apples with oranges."

Different fields of research exhibit quite different citation rates or averages, and the difference can be as much as 10:1. The average 10-year-old paper in molecular biology and genetics may collect 40 citations, whereas the average 10-year-old paper in a computer science journal may garner a relatively modest four citations.

Even within the same field, one should not compare absolute citation counts of an eight-year-old paper with those of a two-year-old paper, since the former has had more years to collect citations than the latter.

Likewise, there is little sense in comparing the thick publication dossier of a researcher who has been publishing for 30 years and runs a large laboratory with the handful of recently published papers from a newly minted Ph.D. in the same field.

This is all really no more than common sense. Still, comparing like with like is the "golden rule" of citation analysis.

6) Use relative measures, not just absolute counts.
This rule applies to citation counts rather than to publication counts, since there is very little data collected on average output for a researcher by field and

over time. The problem is attributing papers to unique individuals in order to calculate typical output. Thomson Reuters carries no marker for unique individuals, only unique name forms.

Eugene Garfield has said that he thinks there is no better single indicator of status and peer regard in science than total citations. And his studies have demonstrated the frequent correlation between scientists with the most citations in their field and those who are chosen for the Nobel Prize. But this is a very select and statistically atypical group of researchers, with thousands or tens of thousands of citations. For these few, absolute citation counts are appropriate.

However, most scientists can claim hundreds—not thousands—of citations. As one deals with smaller numbers, it is important not to put too much weight on relatively minor differences in total citations. Again, it should be recognized that the citation totals of a researcher likely reflect the number of papers produced, the field of research, and how many years the papers have had to collect citations.

Using relative measures. To begin to make distinctions among individuals with a normal, or more typical, number of citations, obtain, among other measures the following:

Absolute counts:
- papers in Thomson Reuters-indexed journals
- papers per year on average
- papers in top journals (various definitions)
- number of total citations

Relative measures:
- citations per paper compared with citations per paper in the field over the same period
- citations vs. expected (baseline) citations
- percent papers cited vs. uncited compared to field average
- rank within field or among peer group by papers, citations, or citations per paper

Field averages are usually generated using journal sets that serve to define the field. This is not always optimal, and everyone has a different idea about such journal-to-field schemes. Although imperfect, such field definitions offer the advantage of uniformity and of measuring all within the same arena, although it is an arena with fuzzy outlines.

Expected or baseline citations are geared to a specific journal, a specific year, and a specific article type (such as review, note, meeting abstract, or letter.) The expected citation score is an attempt to gauge relative impact as precisely as possible based on three combined factors:

1. The year the paper was published, since, as mentioned, older papers have had more time to collect citations than younger ones.
2. The journal in which the paper appeared, since different fields exhibit different average citation rates and, even in the same field, there are high- and low-impact titles.
3. The type of article, since articles and reviews are typically cited more than meeting abstracts, corrections, and letters.

The goal in arriving at the expected citation score is to come as close as possible to the peers for the paper under review—so that like is compared to like.

This is an effective measure for assessing a paper, multiple papers by a researcher, those of a team, and even those of an entire institution.

To examine the publication record of an individual, sum all the actual citation counts to the papers and then sum all the expected citation scores for each paper. The next step is to make a ratio of the two to gauge better than average, average, or lower than average performance, and by how much.

In this methodology, it is as if an exact double (a "doppelganger") of the researcher under review is created. This double would be the ideal of the average researcher. Every time the real researcher published a paper, the double would publish one in the same journal, in the same year, and of the same article type. The papers of the double would always achieve the exact average in terms of citations for such papers. The real researcher would not, of course, but the comparison of the two is often enlightening.

A caution: there are instances where comparing actual to expected citations may mislead. Consider the case of a very prominent, well-regarded biomedical researcher in the United States, one of the elite group of Howard Hughes Medical Institute Investigators. When calculated, his actual-to-expected-citation ratio was a very modest +3 percent, or 3 percent more than average. However, he published almost exclusively in *Science, Nature,* and *Cell* — a very high bar to jump over at every attempt. Needless to say, his total citation score put things in truer perspective.

7) Obtain multiple measures.

The use of multiple measures is a kind of insurance policy against drawing false conclusions from one or two measures alone. The case of the Howard Hughes Investigator demonstrates how using single measures can be misleading. When a variety of different output and impact statistics are collected, whether the subject in question is individuals, groups, or institutions, the multiple measures form a kind of mosaic that describes the influence of the research.

8) Recognize the skewed nature of citation data.

Whether one is reviewing the papers of an individual researcher, those of a research team, papers in a single journal or in group of journals, those of a specific field in a given year, or those of an entire university, lab, or industrial firm, the citation distribution of the dataset will be highly skewed. That is, a small number of papers in the population will be highly cited and the large majority will be cited little or not at all.

This is the nature of these data at every level of analysis. Even the papers of highly cited scientists and Nobel laureates exhibit this type of distribution. It is a rich area of study that has yet to be fully mined or explained, but such distributions—which are anything but so-called normal bell curve distributions—seem to be common to cases where human choice is at play.

A related phenomenon is the concept of criticality. At some critical point, a paper achieves enough citations that other citations to it seem to accelerate. Then

the paper, in citation terms, takes off on a different and higher trajectory. (The challenge of determining this point, as yet to be defined mathematically with a single equation, has begun to attract the attention of theoretical physicists.)

9) Confirm that the data collected are relevant to the question.

10) Ask whether the results are reasonable.

Bibliometricians and those who use these data for science policy and research funding decisions should do no less than follow scientific process for evaluating the data yielded.

- Thus, these two rules, although they represent two different steps, are important to follow together. In doing so, the user will double-check the data collected and view it with the same scientific skepticism with which all data should be viewed: Are the data relevant to the question one originally set out to answer? Can the conclusions from the data be refuted? Are the conclusions beyond the limits of the data collected?

The Overarching Rule of Bibliometrics

It is, of course, the business of Thomson Reuters to advocate the use of quantitative analysis for research evaluation. But we speak out about naïve methodologies and the misleading uses to which Thomson Reuters and others' similar data are sometimes put. The consequences of such misuse can be profound—for individuals, research groups, institutions, journal publishers, and even nations and their national research programs.

The goal of bibliometrics is to discover something, to obtain a better, more complete understanding of what is actually taking place in research. This deeper understanding can better inform those charged with making difficult choices about allocating resources, generally in the context of peer review.

The overarching rule in using bibliometrics, then, is that the results should be presented openly and honestly, that the analysis should be straightforward in its methodology and simple to explain, so that others can understand and check it. Such transparency will help ensure its appropriate use.

TEN RULES IN USING PUBLICATION AND CITATION ANALYSIS

1. Consider whether available data can address the question.
2. Choose publication types, field definitions, and years of data.
3. Decide on whole or fractional counting.
4. Judge whether data require editing to remove "artifacts".
5. Compare like with like.
6. Use relative measures, not just absolute counts.
7. Obtain multiple measures.
8. Recognize the skewed nature of citation data.
9. Confirm that the data collected are relevant to the question.
10. Ask whether the results are reasonable.

And, above all, present the results openly and honestly.

Thomson Reuters
Regional Head Offices

Americas
Phone: +1 800 336 4474
+1 215 386 0100

Europe, Middle East and Africa
Phone: +44 20 7433 4000

Japan
Phone: +81 3 5218 6500

Asia Pacific
Phone: +65 6879 4118

Thomson Reuters Offices around the World
Sydney, Australia
Rio de Janeiro, Brazil
Paris, France
Munich, Germany
Hong Kong
Bangalore, India
Tokyo, Japan
Mexico City, Mexico
Beijing, People's Republic of China
Seoul, Republic of Korea
Singapore
Barcelona, Spain
Taipei, Taiwan
London, United Kingdom
USA
Alexandria, Virginia
Ann Arbor, Michigan
Carlsbad, California
San Jose, California
East Haven, Connecticut
Lisle, Illinois
Portland, Maine
Horsham, Pennsylvania
Philadelphia, Pennsylvania

thomsonreuters.com

AG 05 8061-02 C WP

THOMSON REUTERS